CN 53-1189/P ISSN 1672-7673

R 语言应用于 LAMOST 光谱分析初探*

陈淑鑫[1,2]，罗阿理[3]，孙伟民[2]

(1. 齐齐哈尔大学机电工程学院，黑龙江 齐齐哈尔 161006；2. 哈尔滨工程大学理学院纤维集成光学教育部重点实验室，黑龙江 哈尔滨 150006；3. 中国科学院光学天文重点实验室，北京 100012)

摘要：以可扩展性极强的开源软件 R 程序语言为工具，发挥在统计学和数据挖掘领域强大的数据分析能力，重点研究 R 语言用于读写 FITS 格式文件软件包 RFITSIO 的主要功能和特点，并对 LOMAST 采集的 FITS 文件进行详细介绍，将海量 LOMAST 巡天光谱 DR2 数据用 RFITSIO 读出恒星光谱，并利用 R 语言的主成分分析工具提取各类型光谱数据的特征量即主成分。从含有大量冗余信息的光谱中提取代表恒星光谱特征的主要成分，通过采用主成分分析方法提取光谱特征，重构后能够有效降低原始光谱数据受噪声的影响，为后续数据挖掘工作提供研究基础。

关键词：R 语言；FITSIO；光谱巡天；LAMOST；主成分分析

中图分类号：P114.1 **文献标识码**：A **文章编号**：1672-7673(2017)03-0363-06

R 程序语言是集成了多种数据分析和可视化方法的开源软件，已成为信息时代大数据分析的重要工具，并在统计学和数据挖掘领域可扩展性极强，具备强大的数据分析能力，能有效地简化数据分析的过程。为了将 R 语言的优势应用到天文数据分析中，首先需要了解天文数据的格式。国际通用天文数据格式的标准是普适图像传输系统[1](Flexible Image Transport System，FITS)，该格式由文[1]于 1979 年首先提出，用于描述天文学数据定义和数据本身编码的格式方法，已成为天文学领域应用最广泛的数据格式。前人已将 FITSIO 软件包应用于 FORTRAN 语言、C 语言、IDL 语言中，编写读取 FITS 文件，分析天文数据读取等大量的相关工作。文[2-3]概述了 FITS 文件的标准数据格式以及 3 种类型即图像文件、ASCII 表文件和二进制表文件等，文[2]对 FITS 基本格式及其扩展进行了较详细的分析阐述，文[3]对 g、r、i 3 个波段的 FITS 图像文件进行了重新读写，分析了斯隆数字巡天的测光数据，文[4]利用数据库保存和管理 FITS 头信息归档入库，提供光谱数据文件的查询检索。近年来，由于 R 语言在各专业的统计领域广泛应用，已经有美国马里兰大学 Andrew Harris 等①用 R 语言编写了读取 FITS 的软件包 RFITSIO。

国内外已相继开展了多个天文大规模巡天项目[5]，如 2dF、6dF、RAVE、SDSS、LAMOST 和 Gaia 等，这些大规模光学光谱巡天已获取数以十万、百万甚至千万计的天体光谱[6]。2013 年我国自主研发的大天区面积多目标光纤光谱望远镜[7](Large Sky Area Multi-Object Fiber Spectroscopy Telescope，LAMOST)是目前世界上口径最大、视场范围广、观测目标最多、天体光谱获取率最高的光纤光谱望远镜，获取的 FITS 格式光谱大数据信息达 10^7 数量级。LAMOST 目前已经对国际公开发布了巡天数据约 380 万条，其中未能被软件识别的有 306 810 条。由于大数据导入时包含了大量的冗余信息，所以恒星光谱的全频谱信号充分利用多元统计分析(Multivariable Statistical Analysis)这些数据，本文利用 R 语言提供的数据挖掘工具，通过主成分分析方法消除流量特征之间的相关性，从原始光谱数据中提取与物理参数相关的特征，降维数据并剔除噪声。

* 基金项目：国家自然科学基金(U1631239)；黑龙江省教育厅基本科研业务专项(135109219)；齐齐哈尔市科学技术计划工业攻关项目(GYGG-201518)；齐齐哈尔大学教育科学研究项目(2016072)资助.
收稿日期：2016-11-28；修订日期：2016-12-16
作者简介：陈淑鑫，女，副教授. 研究方向：光谱数据处理. Email：shuxinfriend@126.com
通讯作者：孙伟民，男，教授. 研究方向：天文光子学. Email：sunweimin@hrbeu.edu.cn

① http://www.astro.umd.edu/~harris/r/index.html

1 LAMOST 光谱数据 FITS 格式

20 世纪 80 年代，国际天文联合会正式公布 FITS 格式作为天文数据的国际标准，世界各地的数据中心和天文学家以此文件为标准，完成了相关研究的保存和数据交换。LAMOST 中 FITS 文件已发布的文件名格式为 spec-MMMMM-YYYY_spXX-FFF.fits，扩展名.fits，其中，MMMMM 代表当地修正的儒略日(MJD，即公元前 4713 年 1 月 1 日起计)，YYYY 是计划标识的字符串(PLANID，即计划号)，XX 表示光谱仪的数字编号(在 1 到 16 之间)，并且 FFF 显示了采集光谱的光纤编号(在 1 到 255 之间)。此外，LAMOST 还通过设计关键字“LAMOST JHHMMSS.ss+DDMMSS.ss”的形式指定对应一个目标文件，其中 HHMMSS.ss 是以时分秒为单位的赤经值，DDMMSS.ss 是以度分秒为单位的赤纬值，在 FITS 文件的基本头部单元，可选择符合扩展以及其他可选择的特殊记录，并将头文件分成 8 组包括强制关键字、文件信息关键词、望远镜参数关键字、观测参数关键字、光谱仪参数关键字、天气条件关键字、数据简化参数关键字和光谱分析结果关键字。

2 RFITSIO 软件包简介

20 世纪 70 年代开发的天文数据文件 FITS 格式有别于 GIF、JPG 等图像文件[8]，它作为一种标准的数据格式可同时存储图像和数据表格。20 世纪 90 年代初，美国高能物理科学研究中心[9]研发了功能强大、使用简便的 FITSIO 程序库，在 FORTRAN 程序和 C 语言程序中均可直接调用该程序库，并提供简单的途径完成读、写 FITS 格式的文件。

研究中 R 语言包是从相应的 Comprehensive R Archive Network 镜像站点下载并将其放入库中，加载 FITSIO 包后读取，这个包中包含的功能用于读取 FITS 头文件数据单元(HDUs)的图像和扩展二进制表，以及写入 FITS 图像文件等功能。其中函数 readFITS 能自动识别图像(多维数组)和扩展二进制表、返回数据、头文件和扩展信息列表，函数 readFITS 返回值及参数如表 1。利用 readFrameFromFITS 返回从二进制表头文件数据单元的 R 语言数据框架，这两个函数主要接收头文件数据单元的文件参数、二进制表位等。

表 1 readFITS 返回值及参数列表

Table 1 Return values from readFITS and arguments list

向量	功能
header	将 END 语句输入文件头
hdr	将头文件的各组解析值赋给关键字
imDat	数据数组(图像)
axDat	带轴缩放和标签(图像)的数据帧
col	每个列的数据(二进制表)
hdu	FITS 文件头部和数据单元的位置：“1”即为第 1 个文件头部和数据单元
colNames	列名称向量 TTYPEn FITS 变量
colUnits	列单元的向量 TUNITn FITS 变量
TNULLn	未定义值的定义向量，FITS 变量
TSCALn	缩放倍乘的向量，FITS 变量
TZEROn	零缩放向量，FITS 变量
TDISPn	格式信息向量，FITS 变量

3 R 语言分析光谱数据

R 语言具备强大的统计功能，从分析事物、判别分析对象分类到已知类别数量集的表现特征，聚类分析一定尺度的对象相关性，推断该分析对象事物内在可能存在的规律性。常用的方法有回归分析、判别分析、聚类分析以及探索性分析。天文学中的统计方法从 1996 年以来，如文[10]采用 Fortran 程序、C 语言程序以及 Python 程序，本文实验分析基于 R 语言程序平台。

3.1 数据获取

从 http://dr2.lamost.org/下载 2016 年 6 月 LAMOST 发布的第 2 次巡天共享数据 dr2.7z 共计 156 GB，下载后 DR2 中恒星数据按 CLASS 字段执行分类到恒星 74 个文件夹，实现 FITS 文件的聚类分布及科

学计算，分析输出绘制图表。

3.2　数据处理环境

实验编程及程序运行在 Linux 环境下，采用 R 版本 2.9.0 程序语言编写，载入 FITSIO 软件包。实验下载服务器 https://cran.r-project.org/src/contrib/Archive/FITSio/FITSio_1.2-0.tar.gz，安装命令 install.packages("FITSio")执行后，选择 CRAN mirror 为“32:China(Beijing)”。

3.3　绘制数据一维光谱图

R 语言是动态语言，在编写代码时常采用两种形式：一种形式如同书写文章，将要完成的功能按模块编写后统一运行；另一种形式则是编写一行语句编译该行。RStudio 分为 4 个工作区域，左上区域为“编写代码”。左下区域也可写代码且运行程序数据输出，即能完成编写一句回车编译解释一句的工作区域。右上区域为“工作区及历史记录”，如图 1，使用 RStudio 直观显示文件的环境变量。右下区域包括的主要功能：“Files”查看当前主页下的文件；“Plots”展示运算结果的图案；“Packages”展示系统已安装的软件包，同时勾选载入内存；“Help”查看帮助文档等功能。

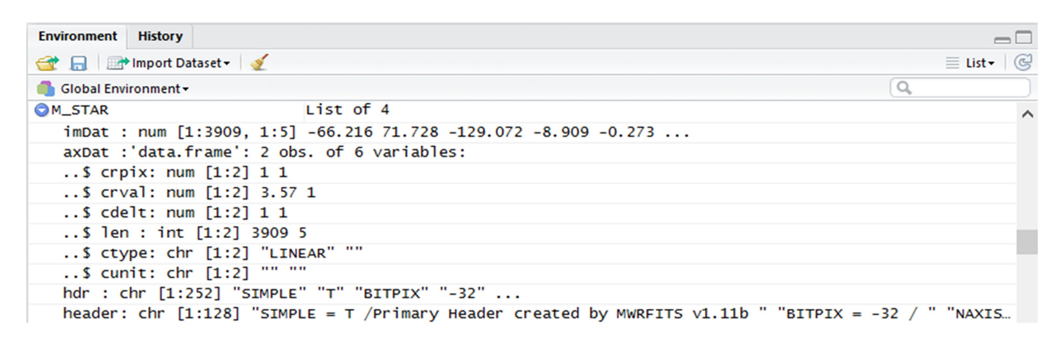


图 1　RStudio 右上工作区读取 M 型星数据变量例表图

Fig. 1　The M Star data variable list in the right top workspace reads of RStudio

LAMOST 发布的巡天光谱 FITS 数据通过 cfitsio 软件包写成，完全依据国际天文联合会发布的天文数据国际标准格式，从已下载的恒星数据中解压 FITS 命名“spec-MMMMM-YYYY_spXX-FFF.fits”格式指定目标文件，利用 require(FITSio)载入需要的 FITSio 软件包，读取光谱数据文件 readFITS("路径*.fits")。利用表 1 中介绍的 readFITS 返回值及参数列表，利用 M_STAR$imDat 提取特征向量矩阵，可将其存储成.csv、.txt 等格式文件，再读取对应参数信息数据列，并限定最大数据范围值。利用 plot()函数选用 type="s"，任意选取 LAMOST 巡天数据库中的 A1IV 型星光谱数据文件 spec-55938-GAC_070N40_V1_sp04-161.fits，绘制流量光谱图如图 2。

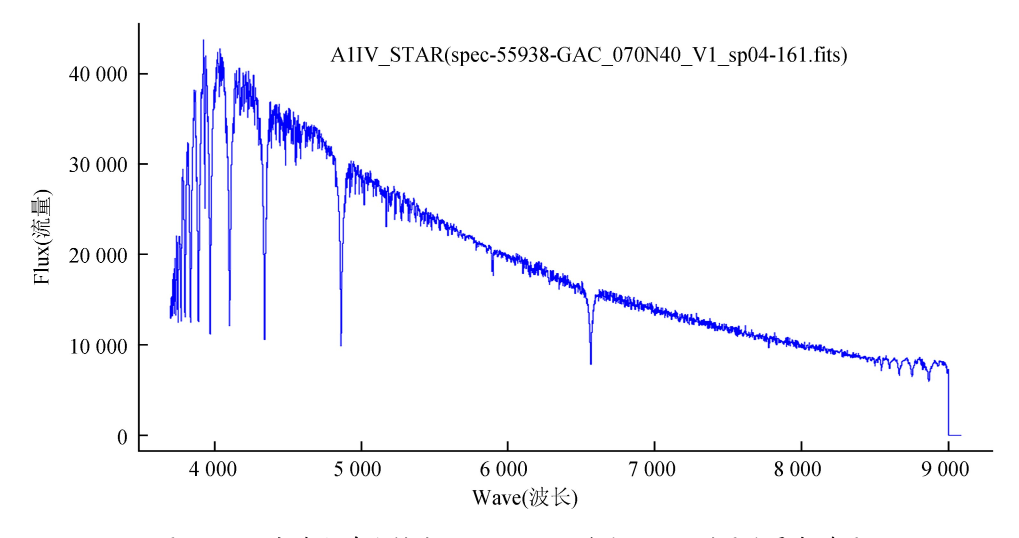


图 2　R 语言读取并绘制的 LAMOST 巡天数据 A1IV 型星流量光谱图

Fig. 2　A spectrum of A1IV type Star in LAMOST survey data read and draw by R

4 挖掘高维光谱数据

实验采用降维数据的方法处理维数无限增大时，简化线性变换掩盖数据原有的信息，探索适合的投影方向，选取最优贡献率将高维空间的目标特征信息尽可能忠实地投影到低维空间，进而高效地提取高维空间中存在的明显差异或特征。

4.1 主成分分析方法应用

高维空间向低维空间线性变换的关键取决于缺失数据的特征信息及关联属性。主成分分析是一种多变量数据线性分析方法，能够较好地完成大样本多参数定量的数据分析，具备非监督性，在尽可能少损失信息的前提下，利用降维理念完成正交变换。由于各主成分变量的总方差贡献率的大小不同，在研究过程中，一般挑选前面几个方差最大的主成分（累计方差贡献率在 80%~90%）分析问题，从而降低问题的复杂程度，抓住主要成分。R 语言在主成分分析中，能高效快捷地对光谱数据进行计算分析。

4.2 R 语言光谱数据主成分分析

为了去除如光子噪声和设备的热噪声等冗余信息的恒星全频谱信号光谱，较好地从原始光谱数据中提取物理参数的相关特征，提高物理参数测量模型的可靠性和运行效率，实验采用 R 语言中核心 stats 包的 prcomp 函数对光谱数据的特征量提取完成主成分分析，主成分分析是基于光谱域的特征提取方法，通过提取光谱中每个波段对应的光谱信号描述能力强弱的流量特征值，旋转坐标系后消除特征之间相关性的光谱分辨力，达到数据降维和剔除噪声的目的。天文研究中主成分分析处理需考虑特征和物理参数间的关系，提取不同的物理参数的特征相关性时难于解释其物理意义，仅仅是数学上的线性相关。现就 LAMOST 大样本、多变量的数据光谱进行主成分分析的应用。

4.3 LAMOST 光谱数据提取主成分分析实验

在 R 语言环境下调用 require(graphics)，运行主成分分析时需去掉提取后的字段名，也可以用 stats 包中的 prcomp 函数及 princomp( ) 函数进行主成分分析，princomp 函数返回一个 princomp 对象，用 summary( ) 函数查询每个主成分的重要信息，用 loadings 函数查看每个变量对主成分的贡献度。此处以 A1IV 类不同光谱数据的主成分为例，分析的结果差别较明显，光谱主成分分析结果如图 3。

```
Importance of components:
                          Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     4.192833e+05 3.315269e+04 9.429371e+03 6.951163e+03
Proportion of Variance 9.909872e-01 6.195700e-03 5.012083e-04 2.723754e-04
Cumulative Proportion  9.909872e-01 9.971829e-01 9.976842e-01 9.979565e-01
```

……

```
                        Comp.769     Comp.770     Comp.771     Comp.772
Standard deviation     1.185081e+01 1.166097e+01 1.161865e+01 1.149543e+01
Proportion of Variance 7.916780e-10 7.665176e-10 7.609638e-10 7.449094e-10
Cumulative Proportion  9.999999e-01 9.999999e-01 1.000000e+00 1.000000e+00
```

图 3 R 语言对 A1IV 型星主成分分析后 summary( ) 的显示结果

Fig. 3 The summary( ) function result of A1IV stellar spectrum Principal Conponent Analysis using in R

根据综合评价函数累计贡献率，然后排序择优，图 3 中 Standard deviation 表示主成分的标准差，对相应特征值的开方，即主成分方差平方根；Proportion of Variance 表示方差贡献率如 Comp. 1 = 99.09%；Cumulative Proportion 表示方差累计贡献率，第 770 个主成分 Comp. 770 时其和为 1。在此例中选取 A1IV 型恒星前 4 个主成分，总方差贡献率达到 99.796%。用这 4 个主成分重构图 3 中 A1IV 型星光谱数据文件 spec-55938-GAC_070N40_V1_sp04-161.fits，重构后如图 4 显示的光谱噪声幅度减小，曲线光滑，而主要的光谱特征谱线都未受损失，说明用主成分分析方法进行特征提取是高效的。

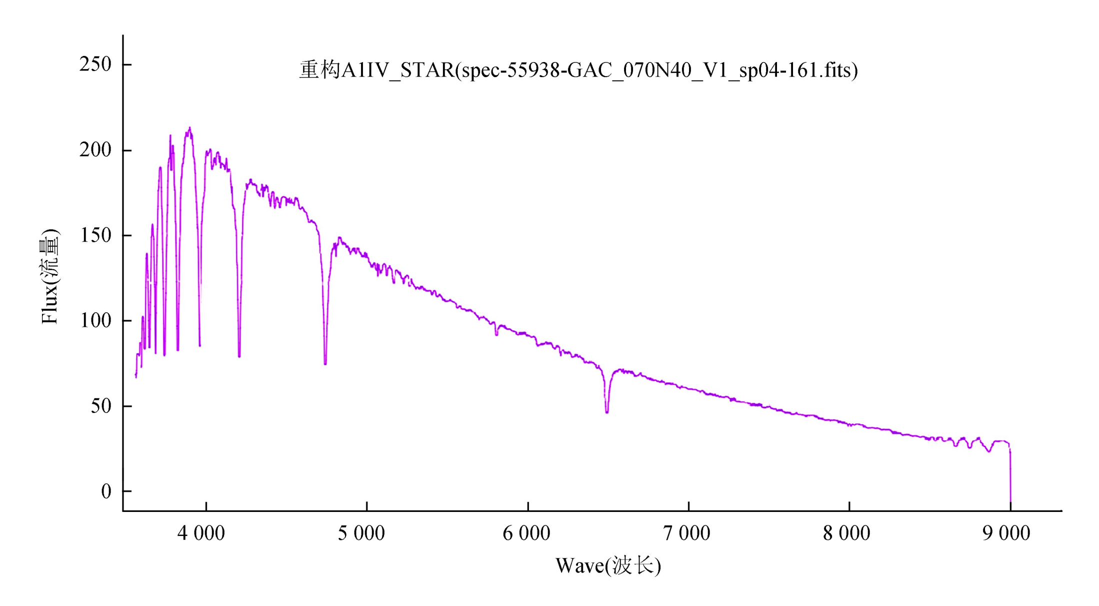


图 4 R 语言重构 A1IV 型星光谱数据文件“spec-55938-GAC_070N40_V1_sp04-161.fits”的流量光谱图

Fig. 4 Reconstraction of a A1IV type spectrum “spec-55938-GAC_070N40_V1_sp04-161.fits” in LAMOST by R

5 结论与展望

目前天文学与云计算在大数据领域的研究应用合作[11]，将逐步提升天文领域更深入地探索宇宙空间，海量数据处理需依靠得力的工具，R 语言平台及其强大的统计分析软件包应用在天文学中将充分发挥其性能优势，在海量数据处理挖掘过程中发挥重要的作用。本文用 RFITIO 对天文数据的读写和主成分分析方法的概念构造光谱的主成分为例，寻找天文光谱数据多变量的“代表”提取到低维空间样本主要特征点，将 R 语言初步应用于天文数据挖掘中，研究结果表明，结合天文数据的领域知识与 R 语言具有的大数据分析优势，能够更加高效地获取挖掘天文光谱数据，是下一步天文数据应用的一项有价值的尝试。

参考文献：

[1] Wells D C, Greisen E W, Harten R H. FITS: a Flexible Image Transport System [J]. Astronomy & Astrophysics Supplement, 1981, 44: 363-370.

[2] 柯大荣，赵永恒. 一种图象传输系统及其 FITS 数据基本格式 [J]. 现代图书情报技术，1994(2): 25-26.
Ke Darong, Zhao Yongheng. An image transport system and its FITS basic format [J]. New Technology of Library and Information Service, 1994(2): 25-26.

[3] 李化南，肖泉宝，邵正义. FITSIO 软件包的简介及应用举例 [J]. 中国科学院上海天文台年刊，2005(1): 119-124.
Li Huanan, Xiao Quanbao, Shao Zhengyi. The introduction to fitsio and its applied example [J]. Annals of Shanghai Observatory Academia Sinica, 2005(1): 119-124.

[4] 崔辰州，李文，于策，等. FITS 数据文件的检索和访问 [J]. 天文研究与技术——国家天文台台刊，2008, 5(2): 116-123.
Cui Chenzhou, Li Wen, Yu Ce, et al. Search and location of FITS data files [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2008, 5(2): 116-123.

[5] 钟守波，韩波，张彦霞，等. 天文大数据管理工具的设计与实现 [J]. 天文研究与技术，2015，12(4)：511-515.
Zhong Shoubo，Han Bo，Zhang Yanxia，et al. Design and implementation of a software tool package for managing massive astronomical data [J]. Astronomical Research & Technology，2015，12(4)：511-515.

[6] 赵永恒. 大规模天文光谱巡天 [J]. 中国科学：物理学 力学 天文学，2014，44(10)：1041-1048.
Zhao Yongheng. Large-scale astronomical spectroscopic surveys [J]. Scientia Sinica：Physica，Mechanica & Astronomica，2014，44(10)：1041-1048.

[7] Luo Ali，Zhao Yongheng，Zhao Gang，et al. The first data release (DR1) of the LAMOST regular survey [J]. Research in Astronomy and Astrohysics，2015，15(8)：1095-1124.

[8] 涂洋，张彦霞，赵永恒，等. 光谱分析软件在天文学研究中的应用 [J]. 天文研究与技术，2016，13(1)：124-132.
Tu Yang，Zhang Yanxia，Zhao Yongheng，et al. Application of spectral analysis softwares in astronomy [J]. Astronomical Research & Technology，2016，13(1)：124-132.

[9] Pence W. CFITSIO，v2.0：a new full-featured data interface [C] // David M Mehringer，Raymond L Plante，Douglas A Roberts. Astronomical Data Analysis Software and Systems VIII，ASP Conference Series，1999，172：487-489.

[10] Babu G J. Feigelson E D. Astrostatistics，Chapman and Hall [C]. (1996) [2016-11-28]. https://www.crcpress.com/Astrostatistics/Babu-Feigelson/p/book/9780412983917.

[11] 崔辰州，于策，肖健，等. 大数据时代的天文学研究 [J]. 科学通报，2015，60(Z1)：445-449.
Cui Chenzhou，Yu Ce，Xiao Jian，et al. Astronomy research in big-data era [J]. Chinese Science Bulletin，2015，60(Z1)：445-449.

Application of R language in LAMOST Spectral Analysis

Chen Shuxin[1,2]，Luo Ali[3]，Sun Weimin[2]

(1. College of Mechanical and Electrical Engineering of Qiqihar University，Qiqihar 161006，China；
2. Key Lab of In-fiber Integrated Optics，Ministry Education of China，Harbin Engineering University，Harbin 150006，China，
Email：sunweimin@hrbeu.edu.cn；3. Key Laboratory of Optical Astronomy，Chinese Academy of Sciences，Beijing 100012，China)

Abstract：The data mining research of large-scale survey is focused on handling，processing and extracting information from massive astronomical data. In this paper，we try to apply the extensible R programming language in LAMOST spectral analysis，and make full use of its capability of integrated data analysis and visualization methods. We mainly study the functions and characteristics of the RFITSIO package for reading and writing FITS format files in R. We then group the LAMOST DR2 data according to the released classification result，and the PCA package in R is applied in each group to extract spectral features from the large amount of noisy spectra. The result shows that，the spectral features are well kept through PCA reconstruction. By extracting the FLUX eigenvalues of the spectral signal description capability of each band in the spectrum，the PCA is used to extract the characteristic value of LAMOST. Rotating coordinate system to eliminate the correlation between the characteristics of the spectral resolution of the data，to reduce the dimensionality of data and remove the effect of noise. This dimensional reduction based feature extraction method can be a very efficient pre-processing approach for the follow-up data mining in LAMOST dataset.

Key words：R language；Flexible Image Transport System Input Output；Spectroscopic Survey；Large Sky Area Multi-Object Fiber Spectroscopy Telescope；Principal Component Analysis